Available online at www.sciencedirect.com

ScienceDirect

International Journal of Approximate Reasoning
49 (2008) 272–284

INTERNATIONAL JOURNAL OF APPROXIMATE REASONING

www.elsevier.com/locate/ijar

Probabilistic approach to rough sets

Wojciech Ziarko

Department of Computer Science, University of Regina, Regina, SK, Canada S4S 0A2

Received 18 April 2006; received in revised form 28 May 2007; accepted 1 June 2007
Available online 22 October 2007

Abstract

The article introduces the basic ideas and investigates the probabilistic version of rough set theory. It relies on both classification knowledge and probabilistic knowledge in analysis of rules and attributes. Rough approximation evaluative measures and one-way and two-way inter-set dependency measures are proposed and adopted to probabilistic rule evaluation. A new probabilistic dependency measure for attributes is also introduced and proven to have the monotonicity property. This property makes it possible for the measure to be used to optimize and evaluate attribute-based representations through computation of probabilistic measures of attribute reduct, core and significance factors.

Keywords: Rough sets; Probabilistic rough sets; Data dependencies; Data mining; Machine learning; Data reduction

1. Introduction

The rough set theory introduced by Pawlak [4] is concerned with finite universes and finite set cardinality-based evaluative measures. It lays out the foundations of the inspiring idea of classification knowledge, in the form of the approximation space, and of the notion of rough set and its approximations. The original theory of rough sets relies on finite universes and on having descriptions of all objects of the universe of interest. In applications, this rarely happens. Typical application scenario involves a partially known universe, represented by a set of samples, based on which rough set analysis is performed. The results are then considered to apply to the whole universe. This kind of approach is common in probabilistic reasoning, with the probability function used to represent relations among sets (events). The probability function values can be estimated from different sources, including assumed distribution functions and set frequencies in a sample. The set frequency estimators of probability theory correspond to set cardinality-based evaluative measures of rough set theory. This correspondence was observed quite early in the development of rough set methodology, leading to a succession of probabilistic generalizations [5–9,11,12,16–19,21–28,30–32] of the original rough set theory. However, the rough set theory methodologies provide additional instruments, originally not present in the probability theory, which allow for deeper analysis of experimental data and for constructing adaptive models of the relations existing in the universe of interest. The probability theory, on the other hand,

E-mail address: ziarko@sasktel.net

0888-613X/$ - see front matter
doi:10.1016/j.ijar.2007.06.014

contributes the basic notion of probability and its estimation, probability distribution evaluative measures, the notion of probabilistic independence and Bayes's equations, which together help to enhance the rough set theory to make it more applicable to real-life problems.

In what follows, the probabilistic version of rough set theory is presented and investigated, partially based on prior results of related research [5–7,9]. In the presentation, clear distinction is being made between classification knowledge and probabilistic knowledge. These two kinds of knowledge are defined in Section 2. The probabilistic notion of event independence is generalized in Section 3, to introduce one-way and two-way measures of set dependencies. One of the measures, the absolute certainty gain, is adopted as a probabilistic rule evaluative parameter and also used for the purpose of rough approximation evaluation. The probabilistic rules, their evaluation and their computation are discussed in Section 4. In Section 5, the variable precision model of rough sets (VPRSM) [6] and the Bayesian rough set model [7,5] are discussed along with some new evaluation measures. The investigation of probabilistic attribute dependencies is presented in Section 6. In particular, the monotonicity of the introduced probabilistic attribute dependency measure, called *λ-dependency*, is proven. This leads to the definition of probabilistic reduct, core and significance factors for attributes, which generalize into probabilistic domain the original notions of rough set theory as introduced by Pawlak [4].

2. Knowledge about universe

In this sections, two kinds of knowledge about universe of interest, the classification knowledge and probabilistic knowledge, involved in the development of the probabilistic approach to rough sets, are discussed.

2.1. Classification knowledge

The rough set approaches are developed within the context of a universe of objects of interest U such as, for example, the collection of patients, sounds, web pages, etc. We will assume here that the universe is infinite in general, but that we have access to a finite sample subset of objects $S \subseteq U$ expressed by accumulated observations about objects in S. The sample represents available information about the universe U. In addition, we will say that a subset $X \subseteq U$ *occurred* if $X \cap S \neq \emptyset$, where $X \cap S$ is a *set of occurrences* of X.

We will also assume the knowledge of an equivalence relation, called the *indiscernibility relation* on U [4], $IND \subseteq U \otimes U$ with finite number of equivalence classes called *elementary sets*. The pair (U, IND) is called the *approximation space*. The collection of elementary sets will be denoted by IND^*. The ability to form elementary sets reflects our *classification knowledge* about the universe U. In the context of this article, the classification knowledge means that each elementary set E is assigned a *description*, denoted as $des(E)$, which specifies a criterion distinguishing all elements of E from its complement. That is, $E = \{e \in U : des(e) = des(E)\}$. The description is usually formed by sets of attribute-value pairs of object properties (attributes), e.g. Age = 50, Gender = female, etc.

Any subset $X \subseteq U$ expressible as a union of some elementary sets is said to be *definable*. Otherwise, the set X is *undefinable*, or *rough* [4]. Any non-elementary definable set will be called a *composed set*. All definable sets have precise descriptions specifying criteria of set membership, whereas rough sets do not. In particular, the descriptions of composed sets are unions of descriptions of their component elementary sets. The classification knowledge is said to be *trivial* (and useless), if there is only one elementary set, corresponding to the whole universe U. The classification knowledge, in the framework of rough set theory, is normally used in the analysis of a *target set* $X \subseteq U$. The target set is usually undefinable. Typical objective of the rough set analysis is to form an approximate definition of the target set in terms of some definable sets.

2.2. Probabilistic knowledge

In probabilistic approaches to rough sets model [7], the classification knowledge is assumed to be supplemented with the *probabilistic knowledge*. The probabilistic knowledge reflects the relative occurrence frequencies of sets (events). It is normally represented by the probability function P defined on σ-algebra of measurable subsets of U. It is assumed here that all subsets $X \subseteq U$ under consideration are measurable by

a probabilistic measure function P with $0 < P(X) < 1$. That is, they are likely to occur but their occurrence is not certain. The probabilistic knowledge consists of three parts:

- For each equivalence class E of the relation IND, it is assumed that its probabilistic measure $P(E)$, that is its relative "size", is known.
- We assume that the conditional probability $P(X|E)$ of X, for each elementary set E, is also known.
- The *prior probability* $P(X)$ of the target set X is known.

Alternatively, the probabilistic knowledge can be expressed equivalently by specifying:

- The *joint* probabilities $P(E \cap X)$ of the *atomic* sets of the form $E \cap X$.
- The prior probability $P(X)$.

The probabilities of elementary sets and the conditional probabilities can be easily calculated from the joint probabilities respectively by

$$P(E) = \sum_{G \subseteq E} P(G) \tag{1}$$

and

$$P(X|E) = \frac{\sum_{G \subseteq X \cap E} P(G)}{\sum_{G \subseteq E} P(G)}, \tag{2}$$

where G denotes and atomic set.

Typically, the probabilities $P(E)$ of elementary sets are estimated based on data by $P(E) \simeq \frac{card(E \cap S)}{card(S)}$, where *card* denotes set cardinality. Similarly, the conditional probabilities $P(X|E)$ can be estimated based on data by approximating the relative degree of overlap between sets X and E by $P(X|E) \simeq \frac{card(E \cap S \cap X)}{card(E \cap S)}$.

3. Probabilistic dependencies between sets

In the presence of probabilistic knowledge, it is possible to evaluate the degree of dependencies between measurable subsets of the universe U. In what follows, we propose two kinds of measures to evaluate the degree of connection or dependency between any two sets. The measures can be seen as generalizations of the well-known notion of probabilistic independence of random events.

3.1. One-way dependency

The first, *one-way dependency* measure is concerned with quantifying the degree of the one-way relation between arbitrary measurable subsets X and Y of U. For the one-way dependency measure, the use of the asymmetric function called *absolute certainty gain* [29] (*gabs*), is proposed:

$$gabs(X|Y) = |P(X|Y) - P(X)|, \tag{3}$$

where $|*|$ denotes absolute value function.

The one-way dependency represents the degree of change of the probability of occurrence of X as a result of the occurrence of the set Y. This is the reflection of the influence of the new information represented by the occurrence of the set Y on the occurrence of the set X. For example, in the medical domain, the test result represented by Y may increase, or decrease, the chances that a patient has disease X, relative to the general probability $P(X)$ of this disease in the population U.

In an approximation space, if the set Y is definable then absolute certainty gain can be computed directly from the available probabilistic knowledge according to the following:

Proposition 1. *If Y is definable in the approximation space (U, IND), then the absolute certainty gain between sets X and Y is given by*

$$gabs(X|Y) = \frac{\left|\sum_{E\subseteq Y} P(E)P(X|E) - P(X)\sum_{E\subseteq Y} P(E)\right|}{\sum_{E\subseteq Y} P(E)}, \tag{4}$$

where $E \in IND^*$

Proof. The proof follows directly from the definition of conditional probability and from the assumption that the set Y is definable. □

The values of the one-way dependency function fall in the range $0 \leqslant gabs(X|Y) \leqslant \max(P(\neg X), P(X)) < 1$. In addition, let us note that if sets X and Y are independent in probabilistic sense, that is if $P(X \cap Y) = P(X)P(Y)$ then $gabs(X|Y) = 0$. This is consistent with the intuition, according to which if X and Y are independent, then occurrence of Y has no effect on probability of occurrence of X. That is, $P(X|Y) = P(X)$, which leads to $gabs(X|Y) = 0$. We may also note that $gabs(U|Y) = 0$ and $gabs(\phi|Y) = 0$, for any measurable subset Y such that $P(Y) > 0$.

3.2. Two-way dependency

The *two-way dependency* measure is concerned with measuring the degree of the mutual connection between measurable sets X and Y. For the two-way measure, the symmetric function *dabs*, called *absolute dependency gain*, is proposed:

$$dabs(X, Y) = |P(X \cap Y) - P(X)P(Y)|. \tag{5}$$

The absolute dependency gain reflects the degree of probabilistic dependency between sets by quantifying the amount of deviation of $P(X \cap Y)$ from probabilistic independence between sets X and Y, as expressed by the product $P(X)P(Y)$. Similarly, $|P(\neg X \cap Y) - P(\neg X)P(Y)|$ is a degree of deviation of the $\neg X$ from total independence with Y. Since $P(\neg X \cap Y) - P(\neg X)P(Y) = -(P(X \cap Y) - P(X)P(Y))$, both target set X and its complement $\neg X$ are dependent in the same degree with any measurable set Y.

As in the case of one-way dependency, if the set Y is definable then the absolute dependency gain can be computed directly from the available probabilistic knowledge, according to the following:

Proposition 2. *If Y is definable in the approximation space (U, IND), then the absolute dependency gain between sets X and Y is given by*

$$dabs(X, Y) = \left|\sum_{E\subseteq Y} P(E)P(X|E) - P(X)\sum_{E\subseteq Y} P(E)\right|, \tag{6}$$

where $E \in IND^*$.

Proof. The proof follows directly from the definition of conditional probability and from the assumption that the set Y is definable. □

The one-way and two-way dependencies are connected by

$$dabs(X, Y) = P(Y)gabs(X|Y). \tag{7}$$

From the above, it follows that the values of the two-way dependency fall in the range $0 \leqslant dabs(X, Y) \leqslant P(Y)\max(P(\neg X), P(X)) < P(Y) < 1$. Also $0 \leqslant dabs(X, Y) \leqslant P(X)\max(P(\neg Y), P(Y)) < P(X) < 1$ i.e. $0 \leqslant dabs(X, Y) < \min(P(X), P(Y))$. In addition, let us note that if sets X and Y are independent in probabilistic sense, that is if $P(X \cap Y) = P(X)P(Y)$ then $dabs(X, Y) = 0$. We may also note that $dabs(U, Y) = 0$ and $dabs(\phi|Y) = 0$, for any arbitrary subset Y such that $P(Y) > 0$. The justification behind all these properties is the same as in the case of one-way dependency.

4. Probabilistic rules

In this part, we are concerned with the evaluation of probabilistic rules using measures introduced in previous sections. In the context of probabilistic approach to rough set theory, probabilistic rules

[7,10,11,13,20,23,25–27] are formal linguistic expressions representing relationships between subsets of the universe U. For any definable subset Y and an arbitrary subset X of the universe U, the *probabilistic rule* is a statement $des(Y) \rightarrow s(X)$, denoted shortly by $r_{X|Y}$, where $s(X)$ is a string of characters used to refer the set X and $des(Y)$ is a description of the set Y. The set Y is referred to as *rule support set*. As opposed to the description of a set, which defines it, $s(X)$ is just a *reference* to a possibly undefinable set, whose description might be unknown. Since rules of this kind are normally used to determine, or to guess, the membership of an object in the set X if the object belongs to the definable set Y, it does not make much sense dealing with rules in which X is definable (if X were definable, the uncertainty-free determination of an object's membership in X could be done based on X's description). Consequently, we will assume that the conclusion part $s(X)$ of the rule $r_{X|Y}$ corresponds to an undefinable set X.

Traditionally, the probabilistic rules $des(Y) \rightarrow s(X)$ are assigned two probabilistic parameters characterizing the relation between sets X and Y:

- The rule $r_{X|Y}$ *certainty* parameter defined as the conditional probability $cert(r_{X|Y}) = P(X|Y)$.
- The rule $r_{X|Y}$ *generality* (also called *support*) parameter defined as the probability $gen(r_{X|Y}) = P(Y)$.

Certainty and generality parameters can be equivalently replaced by certainty and *strength* measures, where the strength is defined as $str(r_{X|Y}) = P(X \cap Y)$. However, rule certainty and generality, or the certainty and strength, do not completely capture the intuitive perception of rule quality. For example, a rule with high certainty $P(X|Y)$ may not be very useful if the prior probability of X is also high. On the other hand, if the prior probability of X is low, a high certainty rule will represent a significant increase in the ability to predict X. Intuitively, such a rule will be very valuable.

To properly represent the degree of *certainty increase* attributed to a probabilistic rule $r_{X|Y}$, relative to the prior probability $P(X)$, the use of the absolute certainty gain parameter $gabs(r_{X|Y}) = gabs(X|Y)$ is proposed. The absolute certainty gain represents the degree of increase of the certainty of prediction of X, as a result of the occurrence of the set Y. As the absolute certainty gain cannot be derived from certainty and generality parameters, we propose that probabilistic rules be evaluated in terms of the three parameters rather than two: generality (or strength), certainty and, additionally, the certainty gain parameter instead of generality (or strength), certainty only.

Any elementary set $E \in IND^*$ corresponds to an *elementary rule* $des(E) \rightarrow s(X)$. The strength, certainty and the absolute certainty gain of elementary rules can be simply obtained from the available probabilistic knowledge. It was shown in Proposition 1 that the absolute certainty gain can be computed from the probabilities associated with the elementary sets. The following Proposition 3 demonstrates that strength and certainty of any probabilistic rule $des(Y) \rightarrow s(X)$ can also be computed in similar way.

Proposition 3. *The strength and certainty of the rule $r_{X|Y} = des(Y) \rightarrow s(X)$ are respectively given by*

$$str(r_{X|Y}) = P(Y) = \sum_{E \subseteq Y} P(E), \tag{8}$$

$$cert(r_{X|Y}) = P(X|Y) = \frac{\sum_{E \subseteq Y} P(E)P(X|E)}{\sum_{E \subseteq Y} P(E)}. \tag{9}$$

Proof. The proof follows directly from the definition of conditional probability and from the assumption that the set Y is definable. □

The practical implication from Propositions 1 and 3 is that once the basic probabilistic knowledge is estimated from data, there is no need to refer to the data set again to compute any kind of probabilistic rules and attribute dependencies.

5. Probabilistic formulations of rough sets

In this section, we review three probabilistic formulations of rough sets: the variable precision rough sets (VPRSM), symmetric variable precision rough sets and Bayesian rough sets. All these formulations originate

from the VPRSM, with the Bayesian formulation being the most relaxed one, based on the concept of probabilistic independence of events in the boundary area, as opposed to the two other definitions which impose parametric constraints when defining approximation regions.

5.1. Variable precision asymmetric rough set formulation

In the VPRSM the probabilistic knowledge, as represented by the probability estimates associated with elementary sets, is used to construct generalized rough approximations of the target subset $X \subseteq U$. The defining criteria are expressed here in terms of conditional probabilities and of the prior probability $P(X)$ of the target set X. The *certainty control* criteria parameters are used to control degree of required certainty gain in the lower approximations of the set X or its complement $\neg X$.

The first parameter, referred to as the *lower limit* l, satisfying the constraint $0 \leqslant l < P(X) < 1$, represents the highest acceptable degree of the conditional probability $P(X|E)$ to include the elementary set E in the negative region of the set X, i.e. in the positive region of its complement $\neg X$:

$$NEG_l(X) = \cup\{E : P(X|E) \leqslant l\}. \tag{10}$$

The second parameter, referred to as the *upper limit* u, satisfying the constraint $0 < P(X) < u \leqslant 1$, defines the *positive region* of the set X. The upper limit reflects the least acceptable degree of the conditional probability $P(X|E)$ to include the elementary set E in the positive region (lower approximation):

$$POS_u(X) = \cup\{E : P(X|E) \geqslant u\}. \tag{11}$$

The boundary area includes elementary sets that are not sufficiently associated, with both the target set X and its complement $\neg X$. This means that the conditional probability of target set occurrence is less than the required upper limit and is higher than the lower limit:

$$BNR_{l,u}(X) = \cup\{E : l < P(X|E) < u\}. \tag{12}$$

The union of the positive region and of the boundary area forms the upper approximation $UPP_l(X)$ of the target rough set:

$$UPP_l(X) = POS_u(X) \cup BNR_{l,u}(X) = \cup\{E : P(X|E) > l\}. \tag{13}$$

The pair consisting of the upper and lower approximations forms the approximate representation of the set X. The accuracy of the approximate representation can be defined, after Pawlak [4], as the ratio or relative "sizes" of the approximations:

$$ACC_{l,u}(X) = \frac{P(POS_u(X))}{P(UPP_l(X))} = P(POS_u(X)|UPP_l(X)). \tag{14}$$

This probabilistic accuracy measure generalizes the measure introduced in [4], to reach value of 1 for definable target set X, i.e. when $POS_u(X) = UPP_l(X)$ for $u = 1$ and $l = 0$.

5.2. Variable precision symmetric rough set formulation

The special case VPRSM is called *symmetric* if $l = 1 - u$ [6,7]. In the symmetric model, the parametric criteria for positive region of the target set X, and for its complement $\neg X$, are identical. In this case, with the precision control parameter denoted as $\beta = u = 1 - l$, the negative region is defined by

$$NEG_\beta(X) = \cup\{E : P(\neg X|E) \geqslant \beta\}. \tag{15}$$

Similarly, the positive region of the set X, is

$$POS_\beta(X) = \cup\{E : P(X|E) \geqslant \beta\} \tag{16}$$

and the boundary area is

$$BNR_\beta(X) = \cup\{E : 1 - \beta < P(X|E) < \beta\}. \tag{17}$$

The modified version of upper approximation is given by

$$UPP_\beta(X) = \cup\{E : P(X|E) > 1 - \beta\} \tag{18}$$

and the accuracy of target set X representation is

$$ACC_\beta(X) = P(POS_\beta(X)|P(UPP_{1-\beta}(X))). \tag{19}$$

Because $\beta > P(X)$, then both positive and negative regions can be expressed in terms of absolute certainty gain:

$$NEG_\beta(X) = \cup\{E : gabs(\neg X|E) \geqslant \beta - P(X)\}, \tag{20}$$

$$POS_\beta(X) = \cup\{E : gabs(X|E) \geqslant \beta - P(X)\}. \tag{21}$$

Consequently, we can define the positive region $POS(X, \neg X) = NEG(X) \cup POS(X)$ of the classification $(X, \neg X)$ by a single formula as

$$POS_\beta(X, \neg X) = \cup\{E : gabs(X|E) \geqslant \beta - P(X)\}. \tag{22}$$

Clearly, the approximation regions for the asymmetric VPRSM [7] can also be expressed in terms of the absolute gain function. The positive region of the classification $(X, \neg X)$ represents the area of desired absolute certainty gain, as expressed by the parameter β. Based on the positive region, probabilistic rules can be computed using any lower approximation-based techniques [1–3,11,14]. All these rules will satisfy the imposed minimum absolute certainty gain requirement $\beta - P(X)$. Since the boundary area is a definable subset of U where the minimum certainty gain requirement is not satisfied, that is

$$BND_\beta(X, \neg X) = \cup\{E : gabs(X|E) < \beta - P(X)\}, \tag{23}$$

no probabilistic rule computed from $BND(X, \neg X)$ will meet the minimum absolute certainty gain threshold of $\beta - P(X)$.

5.3. Bayesian formulation

The definable area of the universe U characterized by the total lack of relationship to the target set $X \subseteq U$ was identified in [7] as the *absolute boundary* region of the set X. In the absolute boundary region, every elementary set E is probabilistically independent from the set X, i.e. $P(X \cap E) = P(X)P(E)$. The boundary area can be expressed by using of the absolute dependency gain function as the criterion:

$$BND^*(X, \neg X) = \cup\{E : dabs(X|E) = 0\}. \tag{24}$$

The area of the universe characterized by at least some probabilistic connection with the target set X is called the *absolute positive region* of the classification $(X, \neg X)$. It can be expressed as

$$POS^*(X, \neg X) = \cup\{E : dabs(X|E) > 0\}. \tag{25}$$

Because $dabs(X|E) > 0$ is equivalent to $P(X|E) > P(X)$ or $P(X|E) < P(X)$, the *absolute positive region* of the classification $(X, \neg X)$ can be broken down into two approximation regions:

- The absolute positive region of the set X:

$$POS^*(X) = \cup\{E : P(X|E) > P(X)\}. \tag{26}$$

- The absolute negative region of the set X:

$$NEG^*(X) = \cup\{E : P(X|E) < P(X)\}. \tag{27}$$

Based on the above definitions, the upper approximation of X can be defined as

$$UPP^*(X) = \cup\{E : P(X|E) \geqslant P(X)\}. \tag{28}$$

After Pawlak [4], the rough set accuracy of approximation of X is the ratio of lower and upper approximations, which in probabilistic terms can be expressed by the conditional probability:

$$ACC^*(X) = P(POS^*(X)|UPP^*(X)). \tag{29}$$

Another useful measure of the quality of approximation of X is the *average (expected) certainty gain* $Egabs^*$ in the positive area of the target set X. The higher gain indicates a classifier with stronger ability to discriminate the target set. It is given by

$$Egabs^*(X) = \sum_{E \in POS^*(X)} P(E|POS^*(X))(P(X|E) - P(X)). \tag{30}$$

The expected gain is bound by the value of $ACC^*(X)(1 - P(X))$, which leads to the normalized form of the gain:

$$Ngabs^*(X) = \frac{Egabs*(X)}{ACC^*(X)(1 - P(X))}. \tag{31}$$

For the full evaluation of the classifier corresponding to the rough set X, the measures of *expected certainty* and *expected strength* can also be defined. Each of these two measures can be associated with any approximation region of X, and any formulation of rough sets, for instance, for the positive region in the Bayesian formulation the expected certainty is given by

$$Ecert^*(X) = \sum_{E \in POS^*(X)} P(E|POS^*(X))P(X|E). \tag{32}$$

Similarly, the expected strength of an elementary set in the positive region can be defined as

$$Estr^*(X) = \sum_{E \in POS^*(X)} P(E|POS^*(X))P(E). \tag{33}$$

The absolute approximation regions form the basis of the Bayesian rough set model investigated in [7,5]. They are also useful in the analysis of probabilistic dependencies between attributes, as demonstrated in the following sections.

6. Attribute-based classification systems

In this section, the attribute value-based classification systems are investigated. The focus is on probabilistic dependencies between attributes and optimal selection of a subset of attributes.

6.1. Elementary, composed and binary attributes

In many applications, the information about objects is expressed in terms of values of observations or measurements referred to as *features*. For the purpose of rough set-based analysis, the feature values are typically mapped into finite-valued numeric or symbolic domains to form composite mappings referred to as *attributes*. A common kind of mapping is dividing the range of values of a feature into a number of suitably chosen subranges via a discretisation procedure. Formally, an attribute a is a function $a : U \rightarrow a(U) \subseteq V_a$, where V_a is a finite set of values called the *domain* of the attribute a. The size of the domain of an attribute a, denoted as $com(a) = card(V_a)$, will be called a *theoretical complexity* of the attribute. The theoretical complexity reflects the maximum number of values an attribute can take. Each attribute defines a classifications of the universe U into elementary sets corresponding to different values of the attribute. That is, each attribute value $v \in a(U)$, corresponds an elementary set of objects $E_v^a = a^{-1}(v) = \{e \in U : a(e) = v\}$. The elementary sets form a partition of U. The equivalence relation corresponding to this partition will be denoted as IND_a, whereas the collection elementary sets will be denoted as IND_a^*. We will divide the attributes into two categories:

- The initial, given collection of attributes A, elements of which $a \in A$ are referred to as *elementary attributes*.
- The *composed attributes*, which are formed by taking combinations of some elementary attributes.

The values of a composed attribute are combinations of values of component elementary attributes. Each composed attribute is a subset of A. For proper reference between an elementary attribute and its value, we

will assume that composed attributes are ordered. For the sake of consistency, we will also treat elementary attributes a as single-element subsets of A, $\{a\} \subseteq A$, and the empty subset of A, $\{\}$ will be interpreted as a *trivial attribute*, i.e. with only one value corresponding to the whole universe U. In the context of this assumption, both elementary and composed attributes C will be perceived in two ways: as subsets $C \subseteq A$ and also as mappings $C : U \to C(U) \subseteq \otimes_{a\in C} V_a$, where $\otimes$ denotes Cartesian product operator of all domains of attributes in C, the domain of C. The theoretical complexity of a composed attribute is a product of theoretical complexities of all its elementary attribute domains, $com(C) = \prod_{a\in C} com(a)$. The theoretical complexity of a trivial attribute is one. In practical applications, the theoretical complexity estimates our ability to learn from example observations, or the *learnability* of a classification represented by an attribute [29]. High theoretical complexity attributes lead to non-learnable classifications.

The lowest complexity, non-trivial attributes are binary-valued attributes. Every non-trivial attribute can be replaced equivalently by a collection of binary attributes. The binary attributes are defined for each value v of the attribute a, by creating a new attribute a_v such that

$$a_v(e) = \begin{cases} 1, & \text{if } a(e) = v, \\ 0, & \text{if } a(v) \neq v. \end{cases} \tag{34}$$

The composed attribute B_a consisting of the binary attributes is equivalent to the attribute a because it generates the same classification of U as the attribute a, that is, $IND_{B_a} = IND_a$. Using binary elementary attributes has a number of advantages, including the consistency of representation, ease of implementation and increased generality of minimal length rules computed by applying the idea of rough set theory value reduct [4]. Consequently, from now on in this article, we will assume that all elementary attributes are binary. The composed attributes are vectors of binary attributes. The theoretical complexity of a composed attribute containing n binary attributes can be simply calculated as 2^n. Therefore, the number of bits n can be used as an alternative complexity measure.

6.2. Probabilistic dependencies between attributes

The presence of non-trivial classification of the universe may improve the degree of the decision certainty. We will assume in this section that the classification IND^*_C corresponds to a composed, in general, attribute $C \subseteq A$. The degree of improvement can be quantified using the expected value $egabs(X|C)$ of the absolute gain functions assigned elementary rules $r_{X|E}$, $E \in IND^*_C$:

$$egabs(X|C) = \sum_{E\in IND^*_C} P(E) gabs(r_{X|E}). \tag{35}$$

The *expected gain function* defined by (35) measures the average degree of increase of the occurrence probability of X or $\neg X$, relative to its prior probability $P(X)$, as a result of presence of the classification knowledge, as represented by equivalence classes of the indiscernibility relation IND_C and the associated probabilities. The notion of the expected gain function stems from the idea of the *relative gain* function reported in [7].

The expected gain function *egabs* can also be seen as the measure of the degree of probabilistic dependency between classification represented by the relation *IND* and the partition of the universe corresponding to the sets X and $\neg X$. This follows from the following proposition:

Proposition 4. *The expected gain function can be expressed as*

$$egabs(X|C) = \sum_{E\in IND^*_C} |P(X \cap E) - P(X)P(E)| = \sum_{E\in IND^*_C} dabs(X, E). \tag{36}$$

Proof. Since $P(X|E) = \frac{P(X\cap E)}{P(E)}$, the term $P(E)gabs(r_{X|E}) = P(E)|P(X|E) - P(X)|$ of (35) can be written as $|P(X \cap E) - P(X)P(E)|$, which demonstrates (36). □

The measure can be also expressed in the following alternative form:

Proposition 5. *The expected gain function can be expressed as*

$$egabs(X|C) = P(X) \sum_{E \in IND_C^*} gabs(E|X). \tag{37}$$

Proof. The formula (37) follows from the Bayes's equation

$$P(X|E) = \frac{P(E|X)P(X)}{P(E)} \tag{38}$$

and from the following identities:

$$P(E)gabs(X|E) = P(E|X)P(X) - P(E)P(X) = P(X)(P(E|X) - P(E)) \quad \text{and}$$
$$P(E)gabs(X|E) = P(E)P(X) - P(E|X)P(X) = P(X)(P(E) - P(E|X)). \qquad \square$$

For the purpose of normalization of the expected gain function, the following Proposition 6 is useful.

Proposition 6. *The expected gain falls in the range* $0 \leqslant egabs(X|C) \leqslant 0.5$.

Proof. Clearly, $0 \leqslant egabs(X|C)$ and $egabs(X|C) = 0$ if and only if for all $E \in IND_C^*$, $P(X \cap E) = P(X)P(E)$. The maximum value of $egabs(X|C)$ is achievable if X is definable, that is if X is a union of some elementary sets. In this case, based on Proposition 4, we have $egabs(X|C) = \sum_{E \in IND^*} |P(X|E) - P(X)P(E)|$. Because X is definable, $egabs(X|C) = \sum_{E \subseteq X}(1 - P(X))P(E) + \sum_{E \subseteq \neg X} P(X)P(E)$. We also note that $\sum_{E \subseteq X} = X$ and $\sum_{E \subseteq \neg X} = \neg X$, which leads to $egabs(X|C) = 2P(X)(1 - P(X))$. The maximum of $egabs(X|C) = 0.5$ is reached when $P(X) = 0.5$. $\square$

The target set X and the attribute C are *independent* if $egabs(X|C) = 0$. The independence can occur only if $P(X \cap E) = P(X)P(E)$, for all elementary sets $E \in IND_C^*$. That is, for the independence between X, or $\neg X$, and the partition IND_C^* to hold, the set X, or $\neg X$, must be independent with each element of the partition IND_C^*. Conversely, the strongest dependency occurs when X is definable and when $P(X) = 0.5$. This would suggest to use of the λ-dependency function $0 \leqslant \lambda(X|C) \leqslant 1$, defined by

$$\lambda(X|C) = \frac{egabs(X|C)}{2P(X)(1 - P(X))} \tag{39}$$

as a normalized measure of dependency between attribute C and the target classification $(X, \neg X)$. The function $\lambda(X|C) = 1$ only if X is definable in the approximation space (U, IND_C), that is if the dependency is deterministic (functional). In line with our initial assumption of $0 < P(X) < 1$, $\lambda(X|C)$ is undefined for $X = \phi$ and for $X = U$.

Finally, because elementary attributes are binary, the λ-dependency function can be used to evaluate the degree of probabilistic dependency between any composed attribute $C \subseteq A$ and an elementary attribute $a \in A$. The dependency will be denoted as $\lambda(a|C)$. To be consistent with this notation, we will use symbol d to denote the *decision attribute* representing the target classification $(X, \neg X)$.

6.3. Optimization and evaluation of attributes

One of the main advantages of rough set methodology is the ability to perform reduction of features or attributes used to represent objects. The application idea of *reduct*, introduced by Pawlak [4] allows for optimization of representation of classification knowledge by providing a systematic technique for removal of redundant attributes. It turns out that the idea of reduct is also applicable to the optimization of probabilistic knowledge representation [5,15], in particular with respect to the representation of the probabilistic dependency between a composed attribute and a binary attribute. The following theorem demonstrates that the probabilistic dependency measure between attributes is *monotonic*, which means that expanding a composed attribute $C \subset A$ by extra bits would never result in the decrease of dependency $\lambda(d|C)$ with the decision attribute d corresponding to the partition $(X, \neg X)$ of the universe U.

Theorem 7. *λ-dependency is monotonic, that is, for any composed attribute $C \subset A$ and an elementary attribute $a \in A$ the following relation holds*:

$$\lambda(d|C) \leqslant \lambda(d|C \cup \{a\}). \tag{40}$$

Proof. To prove the theorem, it suffices to show that the absolute gain function is monotonic. Let F denote any elementary set of the relation IND_C and let E denote an elementary set of the relation $IND_{C\cup\{a\}}$. We will show that $egabs(X|C) = \sum_F P(F)gabs(X|F) \leqslant \sum_E P(E)gabs(X|E) = egabs(X|C \cup \{a\})$.

First, we note that, based on Bayes's equation (38):

$$P(F)gabs(X|F) = P(X)gabs(F|X). \tag{41}$$

Also, if $F \subseteq NEG^*(X)$ then, we have

$$gabs(F|X) = P(F) - P(F|X) = \sum_{E\subseteq F}(P(E) - P(E|X)), \tag{42}$$

$$\sum_{E\subseteq F}(P(E) - P(E|X)) \leqslant \sum_{E\subseteq F} gabs(E|X). \tag{43}$$

It follows that

$$P(F)gabs(X|F) \leqslant P(X)\sum_{E\subseteq F} gabs(E|X). \tag{44}$$

Similarly, we can demonstrate that if $F \subseteq POS^*(X)$ then

$$P(F)gabs(X|F) \leqslant P(X)\sum_{E\subseteq F} gabs(E|X). \tag{45}$$

Last two inequalities imply that

$$\sum_{E\subseteq F} P(F)gabs(X|F) \leqslant P(X)\sum_F \sum_{E\subseteq F} gabs(E|X) \leqslant \sum_E gabs(E|X), \tag{46}$$

which in conjunction with Proposition 5, completes the proof. □

As a consequence of Theorem 7, the notion of the *probabilistic reduct* of attributes $RED \subseteq C$ can be defined as a minimal subset of attributes preserving the dependency with the decision attribute d. That is, the reduct satisfies the following two properties:

$$\lambda(d|RED) = \lambda(d|C) \tag{47}$$

and for any attribute $a \in RED$:

$$\lambda(d|RED - \{a\}) < \lambda(d|RED). \tag{48}$$

The probabilistic reducts can be computed using any methods available for reduct computation in the framework of the original rough set approach [14]. The reduct provides a method for computing fundamental factors in a probabilistic relationship.

Elementary and composed attributes appearing in a reduct can be evaluated with respect to their contribution to the dependency with the target attribute by adopting the notion of a *significance factor*. The significance factor $sig_{RED}(B)$ of an attribute $B \subseteq A$ represents the relative decrease of the dependency $\lambda(d|RED)$ due to removal of B from the reduct:

$$sig_{RED}(a) = \frac{\lambda(d|RED) - \lambda(d|RED - B)}{\lambda(d|RED)}. \tag{49}$$

Finally, as in the original rough set approach, one can define the *core* set of elementary attributes as the ones which form the intersection of all reducts of C, if the intersection is not empty. After [4], any core attribute $\{a\}$ satisfies the following inequality:

$$\lambda(d|C) > \lambda(d|C - \{a\}), \tag{50}$$

which leads to a simple method of core computation.

7. Conclusion

The article is an attempt to introduce a comprehensive probabilistic version of rough set theory by integrating ideas from Pawlak's classical rough set model, elements of probability theory with its notion of probabilistic independence, the variable precision model of rough sets and the Bayesian model. The novel aspects of the approach include the introduction of measures of inter-set dependencies, based on the notion of absolute certainty gain and probabilistic dependence, the adaptation of the absolute certainty gain to probabilistic rule evaluation, the definition of new evaluative measures for probabilistic rough sets, such as probabilistic accuracy and expected gain measures for rough approximation regions, expected strength and certainty of approximation regions. In addition, the notion of a composed attribute was introduced along with the attribute dependency measure based on the idea of expected gain function and its application to attribute optimization and evaluation.

The presented ideas seem to connect well with the general methodology of rough sets, hopefully leading to new applications and better understanding of fundamental issues of data mining and learning from data.

Acknowledgements

The research reported in this article was supported in part by a research grant awarded to the author by Natural Sciences and Engineering Council of Canada.

References

[1] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, ICS Report 1/91, Warsaw University of Technology.
[2] J. Grzymala-Busse, LERS – a system for learning from examples based on rough sets, Intelligent Decision Support: Handbook of Applications and Advances of Rough Sets Theory, Kluwer, 1991, pp. 3–18.
[3] W. Ziarko, N. Shan, A method for computing all maximally general rules in attribute-value systems, Computational Intelligence 12 (2) (1996) 223–234.
[4] Z. Pawlak, Rough Sets – Theoretical Aspects of Reasoning About Data, Kluwer, 1991.
[5] D. Slezak, W. Ziarko, The investigation of the Bayesian rough set model, International Journal of Approximate Reasoning 40 (1–2) (2005) 81–91.
[6] W. Ziarko, Variable precision rough sets model, Journal of Computer and Systems Sciences 46 (1) (1993) 39–59.
[7] W. Ziarko, Set approximation quality measures in the variable precision rough set model, Soft Computing Systems, Management and Applications, IOS Press, 2001, pp. 442–452.
[8] Y.Y. Yao, S.K.M. Wong, A decision theoretic framework for approximating concepts, International Journal of Man–Machine Studies 37 (1992) 793–809.
[9] S.K.M. Wong, W. Ziarko, Comparison of the probabilistic approximate classification and the fuzzy set model, International Journal for Fuzzy Sets and Systems 21 (1986) 357–362.
[10] S. Tsumoto, Modelling medical diagnostic rules based on rough sets, in: Proceedings of RSCTC'1998, Lecture Notes in AI, 1424, Springer-Verlag, 1998, pp. 475–481.
[11] S.K.M. Wong, W. Ziarko, INFER – an adaptive decision support system based on the probabilistic approximate classification, in: Proceedings of the Sixth International Workshop on Expert Systems and their Applications, Avignon, 1986, pp. 713–725.
[12] S.K.M. Wong, W. Ziarko, Algebraic versus probabilistic independence in decision theory, in: Proceedings of International Symposium on Methodologies for Intelligent Systems, Knoxville, 1986, pp. 207–212.
[13] Z. Pawlak, S.K.M. Wong, W. Ziarko, Rough sets: probabilistic versus deterministic approach, International Journal of Man–Machine Studies 29 (1988) 81–95.
[14] W. Ziarko, Rough set approaches for discovery of rules and attribute dependencies, Handbook of Data Mining and Knowledge Discovery, Oxford University Press, 2002, pp. 328–339.
[15] M. Beynon, The elucidation of an iterative procedure to β-reduct selection in the variable precision rough set model, in: Proceedings of RSCTC'2004, Lecture Notes in AI, vol. 1711, Springer-Verlag, 2004, pp. 412–417.
[16] A. Mieszkowicz, L. Rolka, Remarks on approximation quality in variable precision rough set model, in: Proceedings of RSCTC'2004, Lecture Notes in AI, vol. 1711, Springer-Verlag, 2004, pp. 402–411.
[17] D. Slezak, The rough Bayesian model for distributed decision systems, in: Proceedings of RSCTC'2004, Lecture Notes in AI, vol. 1711, Springer-Verlag, 2004, pp. 384–393.
[18] T. Murai, M. Sanada, M. Kudo, A note on Ziarko's variable precision rough set model in non-monotonic reasoning, in: Proceedings of RSCTC'2004, Lecture Notes in AI, vol. 1711, Springer-Verlag, 2004, pp. 103–108.
[19] L. Wei, W. Zhang, Probabilistic rough sets characterized by fuzzy sets, in: Proceedings of RSFDGrC'2003, Lecture Notes in AI, vol. 2639, Springer-Verlag, 2003, pp. 173–180.

[20] S. Tsumoto, Extracting structure of medical diagnosis: rough set approach, in: Proceedings of RSFDGrC'2003, Lecture Notes in AI, vol. 2639, Springer-Verlag, 2003, pp. 78–88.
[21] M. Beynon, Degree of dependency and quality of classification in the extended variable precision rough set model, in: Proceedings of RSFDGrC'2003, Lecture Notes in AI, vol. 2639, Springer-Verlag, 2003, pp. 287–290.
[22] S. Greco, B. Matarazzo, R. Slowinski, J. Stefanowski, Variable consistency model of dominance-based rough set approach, in: Proceedings of RSCTC'2000, Lecture Notes in AI, vol. 2005, Springer-Verlag, 2000, pp. 170–179.
[23] M. Fernandez-Baizan, C. Perez-Lera, J. Feito-Garcia, A. Almeida, LEM3 algorithm generalization based on stochastic approximation spaces, in: Proceedings of RSCTC'2000, Lecture Notes in AI, vol. 2005, Springer-Verlag, 2000, pp. 286–290.
[24] J. Galvez, F. Diaz, P. Carrion, A. Garicia, An application for knowledge discovery based on a revision of VPRS model, in: Proceedings of RSCTC'2000, Lecture Notes in AI, vol. 2005, Springer-Verlag, 2000, pp. 296–303.
[25] S. Tsumoto, An approach to statistical extension of rule induction, in: Proceedings of RSCTC'2000, Lecture Notes in AI, vol. 2005, Springer-Verlag, 2000, pp. 362–369.
[26] M. Wong, C. Butz, Rough sets for uncertainty reasoning, in: Proceedings of RSCTC'2000, Lecture Notes in AI, vol. 2005, Springer-Verlag, 2000, pp. 511–518.
[27] Y. Yao, Probabilistic approaches to rough sets, Expert Systems 20 (5) (2003) 287–291.
[28] S. Greco, B. Matarazzo, R. Slowinski, Rough membership and Bayesian confirmation measures for parametrized rough sets, in: Proceedings of the 10th RSDGRC'2005, LNAI, vol. 3641, Springer, 2005, pp. 314–324.
[29] W. Ziarko, On learnability of decision tables, in: Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing, Uppsala, Sweden, LNAI, vol. 3066, Springer, 2004, pp. 394–401.
[30] S. Greco, B. Matarazzo, R. Slowinski, Parameterized rough set model using rough membership and bayesian confirmation measures 49 (2008) 285–300.
[31] Y. Yao, Probabilistic rough set approximations 49 (2008) 255–217
[32] G. Xie, J. Zhang, K. Lai, L. Yu, Variable precision rough set for group decision-making: an application 49 (2008) 331–343